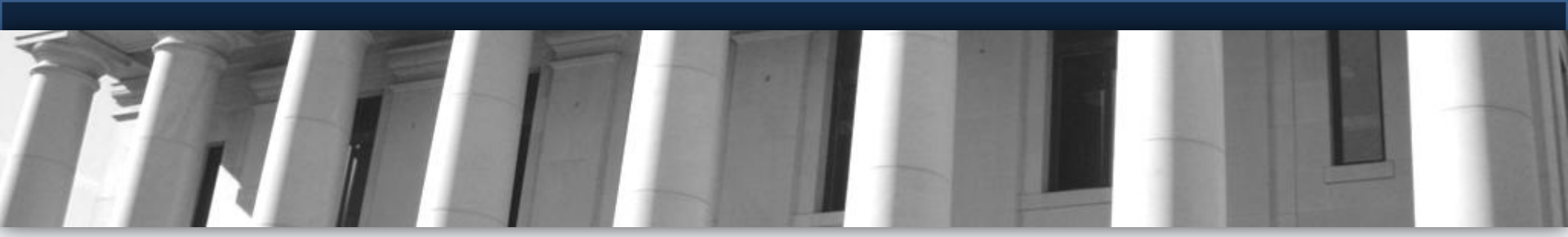

Using Name Change and Non-Education Administrative Data to Assist in Identity Matching

26th Annual Management Information Systems (MIS) Conference

February 14, 2013

Overview

Background

Identity Resolution Challenges

Non-Education Data Sources

How to Apply to Identity Resolution

Value Added

Contact Information

Washington's P20W Data System

- Based in Education Research & Data Center in the state Office of Financial Management
  - Forecasting & Research Division – specialists in education, economics, human services and demography with experience in management and analysis of large administrative data sets
  - Since 1999, home of state's unit-record public baccalaureate enrollment data system
- P20W data system
  - Centralized, research-oriented
  - Comprehensive data from early learning, K-12, public postsecondary, workforce
  - Also apprenticeship, corrections, GED completers, National Student Clearinghouse and selected non-education sources

Washington's P20W Data Warehouse

Informatica HUB

MDM - Master Data Management

ORS - Operational Reference Store

PII DATA with P20_ID Token

All PII data is isolated within the Informatica MDM (Master Data Management) ORS where at P20_ID token is assigned to unique individuals. In addition, a Token_ID is created using a combination of Source System Identifier and Source System Person Identifier and attached to all data received from a system to allow for identity merging and identity unmerging at the P20 Level and at the detailed data level.

P20 Data Warehouse

PERSON
P20_ID
Token

ROLE
ROLE_ID
Token

ORGANIZATION
ORG_ID Token

PRO
P20_ID, ROLE_ID, ORG_ID

PRO Event + Source ID Token

PRO Enrollment + Source ID Token

PRO Achievement + Source ID Token

Names: Challenges in administrative records

Actual name changes – some "official" and some not

- Marriage, Divorce, Adoption
- Personal decision

Different expression of same name

- Use of nicknames
- Missing middle names or middle initial only
- Switched first and middle names
- Cultural name conventions

Universal problems

- High frequency surnames (Smith, Anderson, Nguyen)
- Twins

Some name changes are easy to determine.

Within a single sector:

- **K-12**:

LastName	FirstName	MiddleName	BirthDate	School	K12StateID	SSN
Wilson	John	Edward	1992-12-01	8468	172454	<null>
Anderson	John	Edward	1992-12-01	8468	172454	<null>

- **Postsecondary**:

LastName	FirstName	MiddleName	BirthDate	College	CollegeID	SSN
Smith	Mary	Elizabeth	1990-05-18	365	000392846	532791234
Jones	Mary	Elizabeth	1990-05-18	365	000392846	532791234

- **Workforce** (Unemployment Insurance Wage):

LastName	FirstName	MiddleName	YYYYQ	EmployerID	SSN
Gregg	P	J	20011	A5326B7	533755678
Brown	P	J	20012	A5326B7	533755678

Note: Information presented here has been fabricated to provide illustrative examples. As of June 24, 2011, SSNs beginning with 53279 and 53375 had not been issued by the Social Security Administration.

Cross-sector linking provides resolution

Cross-sector:

- **K-12**:

LastName	FirstName	MiddleName	BirthDate	School	StudentID
Smith	James	Edward	1991-04-06	8468	172454
Smith	Jim	E	1991-06-04	4782	927403
Smith	Bubblegum		1991-06-04	5927	826374

- **Postsecondary**:

LastName	FirstName	MiddleName	BirthDate	College	SSN
Smith	James	E "Bubblegum"	1991-06-04	365	532791234

Note: Information presented here has been fabricated to provide illustrative examples. As of June 24, 2011, SSNs beginning with 53279 had not been issued by the Social Security Administration.

Non-education data source provides resolution

Cross-sector plus additional non-education information:

- **K-12**:

LastName	FirstName	MiddleName	BirthDate	School	StudentID
Smith	James	Edward	1991-04-06	8468	392846
Smith	Jim	E	1991-06-04	4782	927403
Smith	Bubblegum		1991-06-04	5927	826374

- **Postsecondary**:

LastName	FirstName	MiddleName	BirthDate	College	SSN
Smith	James	E "Bubblegum"	1991-06-04	365	532791234

- **Driver license**:

LastName	FirstName	MiddleName	BirthDate	SSN(last 4)
Smith	James	Edward	1991-06-04	1234

(no other James E Smiths – any birthdate – in driver license data)

Note: Information presented here has been fabricated to provide illustrative examples. As of June 24, 2011, SSNs beginning with 53279 had not been issued by the Social Security Administration.

Two people or one?

- **K-12**:

LastName	FirstName	MiddleName	BirthDate	SSN
Anderson	Brittney	Janice	1991-04-06	<null>
Anderson	Brittney	T	1991-04-06	<null>

- **Driver License**

LastName	FirstName	MiddleName	BirthDate	SSN (last 4)
Anderson	Brittney	Janice	1991-04-06	1234
Anderson	Brittney	Theresa	1991-04-06	5678

Note: Information presented here has been fabricated to provide illustrative examples.

First-Middle-Last format doesn't fit all

María Theresa Garcia López (birth date same in all records)

- **K-12**:

LastName	FirstName	MiddleName	School	StudentID
Lopez	Maria	Theresa Garcia	8468	392846
Garcia	Ma Theresa		4782	927403
Lopez	Theresa	Garcia	5927	826374

- **Postsecondary**:

LastName	FirstName	MiddleName	College	SSN
Garcia Lopez	Maria	Theresa	365	532791234
Garcia	Lopez	M	240	532791234

- **Driver License**

LastName	FirstName	MiddleName	SSN (last 4)
Garcia Lopez	Maria Theresa		1234

Note: Information presented here has been fabricated to provide illustrative examples. As of June 24, 2011, SSNs beginning with 53279 had not been issued by the Social Security Administration.

For discussion of cultural naming conventions, see Marcus, N., Adger, C.T., & Arteagoitia, I. (2007). *Registering students from language backgrounds other than English* (Issues & Answers Report, REL 2007-No. 025). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Education Laboratory Appalachia. Retrieved from http://ies.ed.gov/ncee/edlabs.

Name Change Data: Old Names / New Names

Four sources of non-education name change data:

1. WA State court system name changes
2. WA State Department of Licensing data
3. WA State marriage data, for women only
4. WA State divorce data, for women only

With all four sources, raw data is massaged into old name / new name pairs

- For divorce data, the potential old last name is inferred from the husband's last name.

Using Old Name / New Name Pairs

The old name / new name pairs act as a bridge:

- Used to create tuples of data where one name matches an “old name” and the “new name” matches a different name.*
- In practice, an exact match is done on the first and last names only in the tuples.
- Example:
  - $Name_{1A}$ = Joy V. Chuit
  - Old Name/ = Joy Volanda Chuit

    New Name = Roberta S. Almeida
  - $Name_{1B}$ = Roberta Almeida
- Then the resulting data set is organized into “classes” based on similarities in the middle names.

* Subject to the birth dates being the same

Using “classes” to organize potential matches

Potential matches are organized by middle name based classes:

- **Class 1**: The middle names in tuple match perfectly.
  - **Class 1b**: As above, but the day and month of birth is Jan. 1st
- **Class 2**: Somewhere in tuple a full middle name matches a middle initial where only a middle initial is available.
  - **Class 2b**: As above, but the day and month of birth is Jan. 1st
- **Class 3**: Somewhere in tuple, a null middle name matches a non-null middle name
  - **Class 3b**: As above, but the month and day of birth is Jan. 1st

These potential matches are then reviewed in a spreadsheet format.

Value added by use of non-education sources

- Enhances accuracy of longitudinal tracking → more accurate calculation of graduation rates, postsecondary enrollment rates, etc.
  - Reduced undercount of numerators
  - Reduced overcount of denominators
- Reduces bias
  - More complete and accurate information for certain subgroups (name changes after marriage/divorce, blending of families)
  - Improves matching and linking of names from a variety of cultural backgrounds

Contact Us

John Sabel john.sabel@ofm.wa.gov
Carol Jenner carol.jenner@ofm.wa.gov

Washington
Education Research & Data Center www.erdc.wa.gov